Faculty Working Papers

ON THE ESTIMATION OF PRODUCTION FRONTIERS:
Maximum Likelihood Estimation of the Parameters of a Discontinuous Density Function

D. J. Aigner, T. Amemiya, and D. J. Poirier

#236

College of Commerce and Business Administration
University of Illinois at Urbana-Champaign

FACULTY WORKING PAPERS

College of Commerce and Business Administration

University of Illinois at Urbana-Champaign

March 4, 1975

ON THE ESTIMATION OF PRODUCTION FRONTIERS: Maximum Likelihood Estimation of the Parameters of a Discontinuous Density Function

D. J. Aigner, T. Amemiya, and D. J. Poirier

#236

ON THE ESTIMATION OF PRODUCTION FRONTIERS:
Maximum Likelihood Estimation of the Parameters of a Discontinuous Density Function

by

D. J. Aigner, T. Amemiya, and D. J. Poirier*

This paper reports on our attempts to construct a class of estimators for the classical linear regression model that allows for different weights to be placed on positive and negative residuals. The motivation for studying such a problem derives mainly from previous efforts to quantify the notion of a "frontier" production function (Afriat (1972), Aigner and Chu (1968), Timmer (1969, 1971)). Obviously a true frontier involves only one-sided residuals. But there are reasonable objections to the "full" frontier function (Timmer (1971)) that call for some compromise between it and the usual "average" production function. Weighting positive and negative disturbances in a quadratic criterion function offers one way of approaching this compromise.

There is at least one other justification for considering an asymmetric criterion function of the sort just described. It lies in treating asymmetric consequences of under- or over-forecasting in a regression context. But as the reader will note in what follows, we concentrate on within-sample forecasting. Were primary interest focused on asymmetric

---

*University of Wisconsin, Stanford University, and University of Illinois, respectively. Financial support from the National Science Foundation through grants GS-39995 (DJA) and GS-39906 (TA) is gratefully acknowledged. Our thanks go to Roger Koenker for his several contributions and to R. H. Day, A. R. Gallant, A. S. Goldberger, M. J. Hartley and participants in the NBER-NSF Workshop on Segmented and Switching Regressions held at Madison, Wisconsin, June 3-4, 1974 for comments on an initial version. (A previous draft of this paper also appeared under the title, "Regression Estimation with an Asymmetric Criterion Function," SSRI Workshop Paper #7408, University of Wisconsin-Madison, July 1974.) None of them should be held responsible for any weaknesses that remain.

losses of under- or over-forecasting for a particular vector of out-of-sample values for exogenous variables, we would approach the problem in a different manner (cf. Poirier (1973, Ch. 9)). In any event, consideration of the problem as posed allows for a unified treatment of frontier estimation, ordinary least squares, and intermediate cases of apparent empirical interest.

1. The Statistical Model

We assume that a sample of n independent observations are available, having been generated by the model

$$(1) \qquad \underline{y} = \underline{X\beta} + \underline{\varepsilon},$$

where $\underline{y}$ is an $n \times 1$ vector of observations on the dependent variable, $\underline{X}$ is an $n \times k$ matrix of observations on k fixed regressors (including a column of ones), and $\underline{\beta}$ is the $k \times 1$ vector of unknown regression coefficients. Finally, each element of the $n \times 1$ disturbance vector $\underline{\varepsilon}$ is determined by

$$(2) \qquad \varepsilon_i = \begin{cases} \overset{*}{\varepsilon}_i / \sqrt{1-\theta} & \text{if } \overset{*}{\varepsilon}_i > 0 \\ \overset{*}{\varepsilon}_i / \sqrt{\theta} & \text{if } \overset{*}{\varepsilon}_i \leq 0 \end{cases} \qquad i=1, \ldots, n$$

where $\overset{*}{\varepsilon}_i \sim N(0,\sigma^2)$ for $0 < \theta < 1$, and $\overset{*}{\varepsilon}_i$ has either the negative or positive truncated normal distribution (mean $\pm$ .798σ, variance .363σ^2) when $\theta = 1$ or $\theta = 0$, respectively.

The density function thus defined is discontinuous at $\varepsilon_i = 0$. Nevertheless, moments of ε_i exist and are easily derived. For example, using (2) and calculating the appropriate partial moments, we find:

(3) $$E(\varepsilon) = \frac{\sigma}{\sqrt{2\pi}} \left(\frac{\sqrt{\theta} - \sqrt{1-\theta}}{\sqrt{\theta}\ \sqrt{1-\theta}}\right)$$

and

(4) $$V(\varepsilon) = \frac{\sigma^2}{2\theta(1-\theta)} \left\{1 - \frac{(\sqrt{\theta} - \sqrt{1-\theta})^2}{\pi}\right\}, \text{ for } 0 < \theta < 1.$$

Moreover, a likelihood function can be formulated that encompasses these underlying assumptions, and it will be of the form (concentrated over σ^2):

(5) $$\ln L(\underline{y}|\underline{\beta}, \underline{\varepsilon}, \theta) \propto \frac{n_1}{2} \ln \theta + \frac{n_2}{2} \ln (1-\theta)$$

$$- \frac{n}{2} \ln \frac{1}{n} \{\theta \sum_{\varepsilon_i \leq 0} \varepsilon_i^2 + (1-\theta) \sum_{\varepsilon_i > 0} \varepsilon_i^2\}$$

where n_1 is defined to be the number of terms in $\sum_{\varepsilon_i \leq 0}$. From a computational point-of-view the likelihood as stated generally involves k β_j's, θ, <u>and</u> n ε_i's. In effect, we are asked to determine the ε_i's through $\underline{\beta}$ (from the model) and to place each one in the "appropriate" sum (i.e., weight each ε_i^2 by either θ or $(1-\theta)$).

To clarify this latter statement, suppose we define the indicator variables $\{z_i\}$ by

(6) $$z_i = \begin{cases} 1 \text{ if } y_i - \underline{X}_i'\underline{\beta} \leq 0 \\ 0 \text{ if } y_i - \underline{X}_i'\underline{\beta} > 0 \end{cases} \qquad i=1, \ldots, n$$

where $\underline{X}_i'$ is the ith row of X. Then (5) can be rewritten as

(7) $$\ln L(\underline{y}|\underline{\beta}, \underline{z}, \theta) \propto + \Sigma_{i=1}^{n} [z_i \ln \theta + (1-z_i) \ln (1-\theta)]$$

$$- \frac{n}{2} \ln \frac{1}{n} \Sigma_{i=1}^{n} [z_i\theta + (1-z_i)(1-\theta)]\varepsilon_i^2,$$

which is seen to involve k "main" parameters (the β_j's) and n "nuisance" parameters (the z_i's), along with θ. From this formulation it is also apparent that the parameter n_1 in (5) is itself not a free parameter, but is defined as $n_1 = \Sigma_{i=1}^{n} z_i$.

This way of presenting the model makes it reminiscent of the "λ-method" for switching regressions due to Quandt (1972) where $\lambda = \frac{1}{2}$ is known in advance. In fact, the likelihood (5) can be derived within Quandt's framework under the assumption that observations are equally likely to have been generated by a negative or positive half-normal distribution with parameters σ^2/θ or $\sigma^2/1-\theta$, respectively.

The problem of maximizing (5) is not trivial. Moreover, whether the resulting estimators possess all the usual maximum likelihood (ML) properties of consistency, asymptotic normality, and efficiency is not obvious, since the z_i's are discrete variables. Leaving this matter in abeyance for the moment, suppose that merely for computational ease we consider the "minimum distance" estimator of $\underline{\beta}$ (and $\underline{\varepsilon}$) determined by minimizing

$$(8) \qquad S(\underline{y}|\underline{\beta}, \underline{\varepsilon}, \theta) = \theta \sum_{\varepsilon_i \leq 0} \varepsilon_i^2 + (1-\theta) \sum_{\varepsilon_i > 0} \varepsilon_i^2,$$

which is seen to be equivalent to maximizing the second term in (5). This asymmetric criterion function contains ordinary least squares ($\theta = 1/2$) and "full" frontier estimation ($\theta = 0$ or $\theta = 1$) as special cases, and is equivalent to ML estimation in these instances. Otherwise, it allows for unequal weights to be given to disturbances of differing sign,

under the interpretation that a larger relative weight should be given to less variable disturbances.

Some discussion of the above specifications seems appropriate at this point. Though discontinuous, the density function for each ε_i is perfectly legitimate. Moreover, the ε_i's are homoscedastic but with non-zero mean. Therefore, ordinary least squares (OLS) applied to (1) without regard to the actual value of θ will in general produce BLUE and consistent estimators of all coefficients except the intercept. In a production function context, hypothesis tests concerning any element of $\underline{\beta}$ except the intercept parameter can be carried out as usual, based on OLS. Indeed, if the sample is a cross-section of firms in a particular industry, firms can even be appropriately ranked for relative efficiency based on their OLS estimated disturbances, since the biased (and inconsistent) intercept estimator affects them all similarly.[1] Use of the criterion function (8) or the "full" likelihood function (7) is aimed at obtaining consistent and asymptotically efficient estimators of all parameters, including the intercept term, through an "appropriate" weighting of observations.

Interpretation of θ as a measure of relative variability of observations above and below the point $\varepsilon_i = 0$ follows easily from the following scenario. Again within the industry production function context, if the source of (random) difference between firms in their "production" of y for given x derives only from inherent differences in the availability of and/or ability to utilize "best practice" technology, the appropriate error distribution should be one-sided ($\varepsilon_i \leq 0$). If either symmetric measurement error (in y) or the influence of a symmetric and additive random input are considered as well, it is apparent that the relative

[1] Several of these points have been made independently by P. Schmidt (1974) in a recent note.

variability in y will differ for firms above and below the point $\varepsilon_i = 0$. How different is what θ measures, and justifies why, for example, θ might be set equal to one ("full" frontier). As technological differences dominate the aforementioned symmetric error influences, $\theta \to 1$. Otherwise $0 \leq \theta < 1$, reflecting the relative importance of these "error components" in determining the observed distribution of firms.

2. Estimation with θ Known

Since in (8) the index sets for the two summations are endogenously determined, it would appear there will be difficulty in locating the minimum of $S(\underline{y}|\underline{\beta}, \underline{\varepsilon}, \theta)$ even if θ is known. However, it is shown in Appendix A that if a unique global minimum exists for the problem

$$(9) \qquad \min_{\{\underline{\beta},\underline{\varepsilon}\}} \; S(\underline{y}, \theta|\underline{\beta}, \underline{\varepsilon}) = \theta \sum_{\varepsilon_i \leq 0} \varepsilon_i^2 + (1-\theta) \sum_{\varepsilon_i > 0} \varepsilon_i^2$$

$$\text{s.t.} \quad \underline{y} = \underline{X}\,\underline{\beta} + \underline{\varepsilon}$$

then the same solution is obtained for the problem

$$(10) \qquad \min_{\{\underline{\beta}^+,\underline{\beta}^-,\underline{\varepsilon}^+,\underline{\varepsilon}^-\}} \; S^*(\underline{y}, \theta|\underline{\beta}^+, \underline{\beta}^-, \underline{\varepsilon}^+, \underline{\varepsilon}^-) = \theta \sum_{i=1}^{n} (\varepsilon_i^-)^2 + (1-\theta) \sum_{i=1}^{n} (\varepsilon_i^+)^2$$

$$\text{s.t.} \quad \underline{y} = \underline{X}\,\underline{\beta}^+ - \underline{X}\,\underline{\beta}^- + \underline{\varepsilon}^+ - \underline{\varepsilon}^-,$$

$$\underline{\varepsilon}^+ \geq \underline{0}, \quad \underline{\beta}^+ \geq \underline{0}$$

$$\underline{\varepsilon}^- \geq \underline{0}, \quad \underline{\beta}^- \geq \underline{0}$$

with $\underline{\varepsilon}^+$, $\underline{\varepsilon}^-$ being n × 1 vectors of signed disturbances, $\underline{\beta}^+$, $\underline{\beta}^-$ being k × 1 vectors of signed parameters, and where also X'X is assumed to be non-singular. This latter problem is formulated as a quadratic programming (QP) problem in 2n + 2k unknowns.[2]

A proof of the inconsistency of the $\hat{\underline{\beta}} = (\underline{\beta}^+ - \underline{\beta}^-)$ that emerges from a solution to (10) is contained in Appendix B. (But recall that only one element of $\hat{\underline{\beta}}$ is inconsistent: the intercept estimator.) Therefore, our formalization of Timmer's suggestion in terms of this particular asymmetric weighting of residuals leads to the conclusion that nothing is to be gained from the effort, at least insofar as bias correction in the intercept is concerned.[3] Consistent estimators of the intercept are available,

---

[2]A key result in proving that (9) and (10) pose equivalent problems is to show min $(\varepsilon_i^+, \varepsilon_i^-) = 0$ for all i and min $(\beta_j^+, \beta_j^-) = 0$ for all j. This is easily argued by contradiction. (See Appendix A). Effectively, then, there are n+k unknowns in the problem.

Along these same lines, see the interesting programming applications to unbiased estimation in the classical regression model contained in Sielken and Hartley (1973).

[3]This last remarks leads us to point out again that OLS is a special case of (10), when $\theta = 1/2$. No efficiency gains of estimation will be realized by recognizing the signs of residuals in the usual OLS criterion function either, whereas it might seem there is latitude for that to be the case. The point is that *a priori* information on the signs of residuals is not involved, but merely an assignment into index sets which then receive the *same* weight in the criterion function. However obvious this conclusion might appear, we provide a formal demonstration of it in Appendix C.

however.

One technique is "corrected" least squares, using equations (3) and (4). In much the same way as the OLS intercept estimate can be "corrected" for bias in the Cobb-Douglas model with a multiplicative log-normal disturbance, in this problem $E(\varepsilon_i)$ is a function of known and estimable parameters (for known θ) and such correction can also be accomplished. Therefore, if $V(\varepsilon)$ is consistently estimable, say by a statistic $\hat{V}(\varepsilon)$, (4) can be manipulated to yield

$$(11) \qquad \hat{\sigma}^2 = \frac{\hat{V}(\varepsilon)[2\theta(1-\theta)]}{\{1 - \frac{(\sqrt{\theta} - \sqrt{1-\theta})^2}{\pi}\}}$$

and, using (3),

$$(12) \qquad \widehat{E(\varepsilon)} = \frac{\sqrt{\hat{V}(\varepsilon)}\ (\sqrt{\theta} - \sqrt{1-\theta})}{[\pi - (\sqrt{\theta} - \sqrt{1-\theta})^2]^{1/2}}$$

These equations provide consistent estimators for σ^2 and $E(\varepsilon)$, respectively. For example, if $\theta = 1/3$, $\widehat{E(\varepsilon)} = -0.1362\sqrt{\hat{V}(\varepsilon)}$. Therefore, writing the OLS intercept estimator as b_1, the "corrected" OLS intercept estimator is given simply by:

$$(13) \qquad \hat{\beta}_1 = b_1 - \widehat{E(\varepsilon)}$$

which is $\hat{\beta}_1 = b_1 + 0.1362\sqrt{\hat{V}(\varepsilon)}$ for $\theta = 1/3$. Since OLS will yield an unbiased estimator for $V(\varepsilon)$ no matter what the true θ may be, (13) can be implemented (for known θ) and may produce a useful estimator.

Furthermore, it is argued in Appendix B that minimizing the criterion function

(14) $$\tilde{S}(\underline{y}, \theta|\underline{\beta}, \underline{\varepsilon}) = \sqrt{\theta} \sum_{\varepsilon_i \leq 0} \varepsilon_i^2 + \sqrt{1-\theta} \sum_{\varepsilon_i > 0} \varepsilon_i^2$$

with respect to $\underline{\beta}$ (for known θ) will yield consistent estimates of <u>all</u> regression parameters, including the intercept. Computationally this is as difficult a problem as (10). In both cases, (8) and (14), since θ is presumed known at this point, the relevant variance-covariance matrix is available immediately upon recognizing that these minimum distance estimators have the form of generalized least squares estimators given a partitioning of the sample.

Finally, since it is known that the median of the $\underline{\varepsilon_i}$'s is zero, the minimum absolute deviations (MAD) estimator is also consistent for $\underline{\beta}$. Likewise, if we concentrate on consistent estimation of β_1, the intercept, then, for the same reason, the sample median of $\{y_i - \underline{X}'_{2i}\underline{b}_2\}$, where $\underline{X}'_{2i}$ is the ith row of X excluding the first element and $\underline{b}_2$ is the (k-1) × 1 vector of OLS slope estimates, is consistent for β_1.

3. <u>θ Unknown</u>.

The techniques just discussed can be modified to encompass estimation of θ. In order to present those modifications we must first derive the "full" ML estimators for θ and $\underline{\beta}$.

As shown in Appendix B, differentiating (5) with respect to θ and solving the resulting first-order equation gives as the ML estimator, $\hat{\theta}$,

$$\hat{\theta} = \frac{\frac{1}{n_2} \sum_2 (y_i - \underline{X}_i'\underline{\beta})^2}{\frac{1}{n_1} \sum_1 (y_i - \underline{X}_i'\underline{\beta})^2 + \frac{1}{n_2} \sum_2 (y_i - \underline{X}_i'\underline{\beta})^2} \tag{15}$$

where, for ease of notation, we write $\sum_1$ for $\sum_{\varepsilon_i \le 0}$ and $\sum_2$ for $\sum_{\varepsilon_i > 0}$, and $\underline{X}_i'$ is the ith row of X. Concentrating the likelihood further by inserting (15) into (5), we find that the ML estimator for $\underline{\beta}$ minimizes

$$Q = \frac{n_1}{n} \ln \frac{\sum_1 (y_i - \underline{X}_i'\underline{\beta})^2}{n_1} + \frac{n_2}{n} \ln \frac{\sum_2 (y_i - \underline{X}_i'\underline{\beta})^2}{n_2} \tag{16}$$

In Appendix B it is shown that the ML estimator based on (16) is consistent for $\underline{\beta}$. In light of the work of Quandt on the switching regression model, this represents the first proof of consistency for such a model, albeit such a special one. Previously, only Monte Carlo results were available that suggested consistency. Apparently, the fact that the "mixture parameter" is known to us, that all other (regression) parameters are identical in the two regimes, and that the disturbance variances in the regimes are related in a particular way, are sufficient to identify $\underline{\beta}$.

Equation (16) is clearly nonlinear in $\underline{\beta}$ and promises to be more complicated to optimize than either (8) or (14). For this reason we focus our attention on iterative techniques that make use of (15) in conjunction

with the methods of the previous section.

One "pseudo-ML" method uses the OLS slope estimators, $\underline{b}_2$, and an iterated intercept estimate, formed by inserting a consistent estimate of β_1 along with $\underline{b}_2$ into (15), calculating $\hat{\theta}^{(1)}$, inserting $\hat{\theta}^{(1)}$ into (13) to get $\hat{\beta}_1^{(1)}$ and iterating. The final estimates are, say, $\hat{\theta}$, $\hat{\beta}_1$, and $\underline{b}_2$. $\hat{\sigma}^2$ is always available through[4]

$$\hat{\sigma}^2 = \frac{1}{n} [\theta \sum_1 (y_i - \underline{X}_i'\beta)^2 + (1-\theta) \sum_2 (y_i - \underline{X}_i'\beta)^2] \tag{17}$$

with estimates of θ and $\underline{\beta}$ inserted. By an argument in Appendix B, the iterative procedure is expected to be stable.

Another procedure, which re-estimates all elements of $\underline{\beta}$ at each step of the iteration combines (15) with (14). To begin, we insert a consistent estimate of $\underline{\beta}$ into (15), calculating $\hat{\theta}^{(1)}$. This value is then used in (14) and we use a QP procedure to minimize (14) with respect to $\underline{\beta}$, yielding $\hat{\underline{\beta}}^{(1)}$. The process is repeated until convergence is achieved.

Obviously this latter method is the more costly. Whether it has any advantages depends on comparative small sample properties. In fact, we are unable to derive asymptotic standard errors for the "full" ML estimators and so have no analytical norms of comparison even in very large samples. We have some Monte Carlo evidence on these questions to report, however, which is the province of the next section.

---

[4] See Appendix B, equation (B.3).

4. Monte Carlo Results.

Study 1. Since the primary motivation for this study lies in production function "frontier" estimation, the first of two Monte Carlo investigations focuses on such an application. The data were taken from Aigner and Chu (1968) and correspond to an earlier study by Hildebrand and Liu of the primary metals industry.

A constant term ($x_{i1} = 1$, $i = 1, \ldots, n$) and two independent variables (x_{i2} and x_{i3}) were included in the model. All data are state aggregates, with x_{i2} corresponding to the natural logarithm of labor and x_{i3} corresponding to the product of the natural logarithms of lagged (one year) capital and the lagged (one year) ratio of the value of equipment to plant. The dependent variable is the value added for output.[5]

Selecting $\beta_1 = .98$, $\beta_2 = .90$, and $\beta_3 = .03$ as close approximations to the parameter estimates found in the Aigner-Chu article, dependent variable observations were generated for various drawings of ε. Letting $\varepsilon^* \sim NID(0, .6245)$, twenty-eight observations on ε were generated from (2) for $\theta = 1/4$ and $\theta = 1/3$.[6] Two different weighting schemes were then used in posing the quadratic programming problem (10), assuming θ is known at the outset: the first corresponds to the inconsistent minimum distance estimator derived from (8), whereas the second uses square root weights, as in (14), and produces a consistent estimator of all elements of $\underline{\beta}$, including the intercept. (Both yield unbiased estimates of slope coefficients.) Lastly, the "full frontier" case was investigated by letting

---

[5]This model corresponds to that of Aigner and Chu (1968, p. 835, eq. (4.1)) which was used to estimate a "full frontier."

[6]Note that a selection of θ is symmetric on either side of $\theta = 1/2$, so that the cases $\theta = 2/3$ and $\theta = 3/4$ are covered by the results reported for $\theta = 1/3$ and $\theta = 1/4$, respectively, except for the fact that the signs of biases are reversed.

$\theta = 1$, in which event the "negative" truncated normal distribution is used to generate residuals.

In this first set of results, therefore, no iterative procedure is considered. We take θ as known, and wish to compare the minimum distance estimators from (8) and (14), along with OLS and "corrected OLS" for β_1 (from equation (13)).

Summary results on 100 replications of samples with the three different values of θ are reported in Tables 1 to 3. Relative performance in estimating the intercept is probably the most interesting comparison, where it is seen that (up to the accuracy available from 100 replications) the minimum distance estimator (14) has smallest (absolute) bias of all the estimators and smaller root mean square error (RMSE) than the other QP estimator. Though biased downward, the OLS and "corrected OLS" estimators for β_1 compare favorably on the RMSE criterion. In this particular example, the correction to OLS brings the estimated intercept very close to the consistent estimator from (14). Much the same conclusion follows from Tables 2 and 3. Our consistent minimum distance estimator (14) shows a slight improvement over OLS in RMSE. (Recall, all estimators are unbiased for β_2 and β_3).

As a check on the accuracy of our empirical frequency distributions based on 100 replications, we also ran additional samples, up to a total

Table 1

Monte Carlo Results for $\beta_1 = .98$

$n = 28$, 100 Replications

	Mean	Median	Standard Deviation (RMSE)
Case 1 ($\theta = 1/4$)			
Equation (8)	1.11	1.28	2.52 (2.52)
Equation (14)	0.933	0.965	1.70 (1.70)
OLS	0.705	0.966	1.79 (1.80)
"corrected" OLS, Equation (13)	0.930	1.19	1.79 (1.79)
Case 2 ($\theta = 1/3$)			
Equation (8)	1.17	1.36	2.39 (2.40)
Equation (14)	0.947	0.971	1.64 (1.64)
OLS	0.810	0.908	1.67 (1.67)
"corrected" OLS, Equation (13)	0.945	1.043	1.67 (1.67)
Case 3 ($\theta = 1$)			
ML	0.897	0.891	0.479 (0.485)

Table 2

Monte Carlo Results for $\beta_2 = .90$

$n = 28$, 100 Replications

	Mean	Median	Standard Deviation (RMSE)
Case 1 ($\theta = 1/4$)			
Equation (8)	0.912	0.908	0.468 (0.468)
Equation (14)	0.840	0.842	0.297 (0.303)
OLS	0.844	0.834	0.314 (0.315)
Case 2 ($\theta = 1/3$)			
Equation (8)	0.942	0.932	0.449 (0.451)
Equation (14)	0.844	0.844	0.286 (0.291)
OLS	0.846	0.839	0.293 (0.295)
Case 3 ($\theta = 1$)			
ML	0.890	0.902	0.095 (0.0955)

Table 3

Monte Carlo Results for $\beta_3 = .03$

$n = 28$, 100 Replications

	Mean	Median	Standard Deviation (RMSE)
Case 1 ($\theta = 1/4$)			
Equation (8)	0.025	0.009	0.117 (0.117)
Equation (14)	0.047	0.042	0.071 (0.073)
OLS	0.046	0.041	0.076 (0.078)
Case 2 ($\theta = 1/3$)			
Equation (8)	0.018	0.010	0.113 (0.114)
Equation (14)	0.046	0.043	0.069 (0.0708)
OLS	0.045	0.039	0.070 (0.072)
Case 3 ($\theta = 1$)			
ML	0.033	0.032	0.023 (0.0232)

of 300 replications. The results for the minimum distance estimator (14) are presented in Table 4, where it is seen that the empirical frequency distributions are settling down and becoming centered around the true values.[7] It would appear that this estimator may, in fact, yield unbiased estimators of all parameters, although we have not yet been able to prove that conjecture.

All this suggests that a more elaborate set of Monte Carlo results might be welcome, to pinpoint the relative precision of various estimators and to consider additional data alternatives. However, these are expensive to obtain for the QP estimator (14). This is the primary reason we do not attempt to implement the iterative procedure that utilizes (14) and (15).

The evidence presented so far is not really conclusive. But it does suggest that, for the case of θ known, OLS estimators of slope coefficients in conjunction with the "corrected" OLS intercept estimator provide satisfactory estimators when compared to the consistent alternative (but expensive) method from (14).

---

[7] The same is true of the results for our inconsistent QP estimator (8). For example, with 300 replications the means of the empirical frequency distributions for its estimates of β_1, β_2, and β_3 are 1.14, 0.901, and 0.0293, respectively.

Table 4

Additional Results for the Minimum Distance Estimator from (14), $n = 28$, 300 Replications

	$\beta_1 = .98$			$\beta_2 = .90$			$\beta_3 = .03$		
	Mean	Median	Standard Deviation (RMSE)	Mean	Median	Standard Deviation (RMSE)	Mean	Median	Standard Deviation (RMSE)
Case 1 ($\theta = 1/4$)	0.998	1.000	1.638 (1.638)	0.902	0.878	0.283 (0.283)	0.029	0.030	0.0680 (0.0680)
Case 2 ($\theta = 1/3$)	1.006	1.006	1.579 (1.579)	0.902	0.878	0.272 (0.272)	0.029	0.029	0.0650 (0.0650)

In the next section of Monte Carlo results, therefore, we explore further the question of the performance of "corrected OLS" estimation of β_1 in conjunction with estimation of θ through an iterative procedure.

Study 2. In this experiment we abstract from the regression context (and any connection with real-world data), and consider the small-sample behavior of one of the "pseudo-ML" estimators mentioned in section 3, based on a model of the form $y_i = \mu + \varepsilon_i$. In particular, the estimation scheme studied uses (15) and (13) in an iteration that begins by inserting the sample median of the y_i's into (15). The resulting estimate of θ, say $\hat{\theta}^{(1)}$, is substituted into (13) along with the sample mean and standard deviation of the y_i's to yield the first-round estimate of μ, say $\hat{\mu}^{(1)}$. The process is repeated until convergence is obtained, whence the estimate of σ^2 is calculated through (17) with the final-round values for $\hat{\theta}$ and $\hat{\mu}$ inserted.

The results of our experiment are reported in Table 5. 200 replications were used, with $\varepsilon_i^* \sim N(0, 0.5)$, $\mu = 1$, and various values of θ. Sample sizes of 10, 20, 50, and 100 are included. The iteration was stopped when the current values of $\hat{\theta}$ and $\hat{\mu}$ differed by no more than 0.001 and 0.005 from their previous values, respectively, or when the number of iterations reached 50, which happened frequently. In situations where the iteration limit was reached, additional samples were drawn until the required number of replications (of converging cases) was obtained.[8]

---

[8] In every case when the iteration limit was reached, the iteration was in a loop. To some extent looping can be controlled by selection of different error limits for $\hat{\theta}$ and $\hat{\mu}$, but not if a particular pre-set level of accuracy is to be achieved. In any event, we chose to base our main results on converging samples and to investigate the non-converging cases separately.

continued

Table 5

Monte Carlo Results for Iterated Estimators
Based on (15) and (13) in the Model $y_i = \mu + \varepsilon_i$.
Entries are Means over 200 Replications
(RMSE in parentheses), $\mu = 1$, $\sigma^2 = 0.5$.

true θ	estimate of: \ sample size	10	20	50	100
0.5	θ	.501 (.221)	.493 (.180)	.508 (.126)	.501 (.0846)
	μ_2	.989 (.363)	1.01 (.264)	.993 (.173)	1.00 (.133)
	σ^2	.372 (.224)	.424 (.165)	.456 (.102)	.492 (.0679)
0.4	θ	.452 (.220)	.414 (.163)	.407 (.111)	.402 (.0812)
	μ_2	.929 (.358)	.973 (.285)	1.00 (.175)	1.00 (.128)
	σ^2	.368 (.223)	.431 (.169)	.465 (.107)	.486 (.0762)
0.3	θ	.367 (.230)	.326 (.163)	.336 (.117)	.311 (.0712)
	μ_2	.940 (.402)	.926 (.293)	.971 (.194)	.984 (.123)
	σ^2	.390 (.252)	.453 (.176)	.480 (.108)	.488 (.086)
0.2	θ	.293 (.216)	.271 (.168)	.234 (.0938)	.231 (.0689)
	μ_2	.857 (.435)	.927 (.305)	.945 (.189)	.959 (.153)
	σ^2	.452 (.271)	.488 (.205)	.508 (.142)	.517 (.103)
0.1	θ	.228 (.227)	.161 (.126)	.139 (.0760)	.129 (.0558)
	μ	.689 (.682)	.848 (.373)	.883 (.272)	.907 (.209)
	$\sigma2$	.653 (.478)	.598 (.317)	.592 (.239)	.585 (.185)

As we would expect, the results show that μ is estimated with smaller bias and better precision as n increases for given θ and as θ approaches 0.5 for given n. (Recall, as θ moves away from 0.5 the distribution of ε_i is becoming more dispersed.) Similar behavior for our estimator of θ is apparent if we use relative precision as an indicator of efficiency rather than absolute precision. σ^2 seems to be generally underestimated (the exception is $\theta = 0.1$) while θ is generally overestimated for any sample size.

Footnote 10 continued

There seems to be little difference from the converging cases either when the last value is used as the estimate or when we take the average value over the loop. For example, for $n = 20$, $\theta = 0.5$, the averaged $\hat{\theta}$'s over non-converging samples are .487 (.184) and .436 (.182) for the last value and average over the loop, respectively, as compared to .493 (.180) as reported in Table 5.

Below is a table of experimental relative frequencies of occurrence of non-converging samples based on the first 100 replications that went into Table 5. There does seem to be a tendency for the proportion of non-

θ \ n	10	20	50	100
0.5	.34	.36	.39	.46
0.4	.32	.49	.46	.41
0.3	.37	.38	.39	.33
0.2	.32	.38	.31	.29
0.1	.26	.28	.30	.17

converging cases to fall off in a southeasterly direction, but there are many exceptions to that general observation.

5. Conclusions.

After a rather lengthy paper, it is most appropriate to keep this final section brief. To do so, we summarize our results as follows.

Insofar as the practical matters of specifying and implementing an econometric framework within which something other than a "full" frontier function can be estimated are concerned, we have presented such a framework and have evaluated (though incompletely) the properties of a variety of procedures that might be used. In effect, we have produced a "theory of outliers" for the frontier function. The upshot of our efforts, including our own evaluation of the findings, may be criticized as being "much ado about nothing", in that the substance of our recommendations involves nothing more than an "adjustment" of the regression intercept and, after all, who is interested in its value anyway? Two responses are appropriate here: First, for forecasting purposes the intercept is important. Second, along with the intercept adjustment comes explicit "placement" of the function, which, of course, was the goal of this exercise to begin with.

A more attractive criticism would question our approach to the problem as being obtuse. For, after all, why can't one begin with an explicit statement of the error process mentioned after equation (8)? In this "explicit" formulation symmetric measurement error might take the traditional normal form, whereas technological differences among firms may be represented by (e.g.) the negative truncated normal. Our preliminary study of the resulting likelihood function shows it is of the same form as the likelihood function used by Amemiya (1973) in his work on the Tobit model, but with differences in how parameters enter. In this context it is also apparent that an equivalent to our θ can be estimated along with other model para-

meters, and is readily interpreted as an indicator of relative variability.

The admission that there is a more direct means by which to capture the behavioral underpinnings of our problem is not meant to detract either from the interesting theoretical results obtained in Appendix B or the possibility that the estimators discussed here may still be preferable to those derived from this alternative statistical model.

Appendix A: Equivalence of equations (9) and (10).

To prove the equivalence of (10) with the original problem (9) requires showing that at a solution point for (10), (a) min $(\varepsilon_i^+, \varepsilon_i^-) = 0$ for all i and (b) min $(\beta_j^+, \beta_j^-) = 0$ for all j. (Actually, the technique of replacing a variable which is to be unrestricted in sign by the difference of two nonnegative variables is a standard operational procedure in mathematical programming.)

To demonstrate (a), suppose min $(\varepsilon_i^+, \varepsilon_i^-) = \alpha > 0$ for some $i = m$ at a solution point $S^*_{opt} = \min \{\theta \, \Sigma_i(-\varepsilon_i^-)^2 + (1-\theta) \, \Sigma_i \, (\varepsilon_i^+)^2\}$. Since S^* is strictly quasi-convex in the ε-variables and the constraint set is linear, such a solution exists and is unique. Then, consider a "new" solution with coordinates $(\varepsilon_m^+ - \alpha)$ and $(\varepsilon_m^- - \alpha)$ replacing ε_m^+ and ε_m^-, respectively. The constraints are obviously satisfied for this solution and the value of the objective function will be smaller than S^*_{opt}, which contradicts the original supposition. Hence min $(\varepsilon_i^+, \varepsilon_i^-) = 0$ for all i.

As to the second proposition, let min $(\beta_j^+, \beta_j^-) = \gamma > 0$ for some $j = \ell$, again at a solution point S^*_{opt}, and consider replacing β_ℓ^+ and β_ℓ^- by $(\beta_\ell^+ - \gamma)$ and $(\beta_\ell^- - \gamma)$, respectively. Clearly the constraints are satisfied for these modified coordinates, but the value of the objective function is left unchanged. To see this, at an index $i = m$ suppose $\varepsilon_m^+ = 0$ (hence $\varepsilon_m^- \geq 0$). Then, $y_m = X_m'\underline{\beta}^+ - X_m'\underline{\beta}^- - \varepsilon_m^-$, where X_m' is the $m^{\underline{th}}$ row of X, and $(\varepsilon_m^-)^2 = (y_m - X_m'\underline{\beta}^+ + X_m'\underline{\beta}^-)^2$. Replacing β_ℓ^+ and β_ℓ^- by $(\beta_\ell^+ - \gamma)$ and $(\beta_\ell^- - \gamma)$ leaves $(\varepsilon_m^-)^2$ unaffected. Since S^* will be strictly quasi-convex in the β-variables for $X'X$ non-singular, this conclusion contradicts the uniqueness of a solution S^*_{opt}. Since $\underline{\beta}^+, \underline{\beta}^- \geq 0$ we conclude that min $(\beta_j^+, \beta_j^-) = 0$ for all j.

Appendix B: Estimation of the Parameters of a Discontinuous Density Function

In this Appendix we develop the main theoretical results used in sections 2 and 3 of the paper. The argument is couched in terms of a simplified version of the model (1), namely when there is only a column of ones in X and the goal is estimation of the mean of y_i, say μ. Extension to the regression case is apparent.

Suppose a random variable y has the density

$$\text{(B.1)} \quad \begin{cases} \dfrac{\sqrt{\theta_0}}{\sqrt{2\pi}\sigma_0} \exp\left[-\dfrac{\theta_0}{2\sigma_0^2}(y-\mu_0)^2\right] & \text{for } -\infty < y \leq \mu_0 \\[2ex] \dfrac{\sqrt{1-\theta_0}}{\sqrt{2\pi}\sigma_0} \exp\left[-\dfrac{1-\theta_0}{2\sigma_0^2}(y-\mu_0)^2\right] & \text{for } \mu_0 < y < \infty. \end{cases}$$

We want to consider the estimation of μ_0, σ_0, and θ_0 on the basis of n independent observations $\{y_i,\ i = 1,2,\ldots,n\}$ on y.

The log likelihood function of $\{y_i,\ i = 1,2,\ldots,n\}$ is

$$\text{(B.2)} \quad \ln L = -\frac{n}{2}\ln 2\pi + \frac{n_1}{2}\ln\theta + \frac{n_2}{2}\ln(1-\theta) - \frac{n}{2}\ln\sigma^2 - \frac{\theta}{2\sigma^2}\sum_1 (y_i-\mu)^2 - \frac{1-\theta}{2\sigma^2}\sum_2 (y_i-\mu)^2$$

where $\sum_1$ means the summation over $\{i|y_i \leq \mu\}$, and $\sum_2$ means the summation over $\{i|y_i > \mu\}$, and n_1 and n_2 are the numbers of terms in $\sum_1$ and $\sum_2$ respectively. Maximizing (B.2) for σ^2, we obtain

$$\text{(B.3)} \quad \hat{\sigma}^2 = \frac{1}{n}\left[\theta \sum_1 (y_i-\mu)^2 + (1-\theta)\sum_2 (y_i-\mu)^2\right].$$

Inserting (B.3) into (B.2), we obtain the concentrated log likelihood function in θ and μ:

$$\text{(B.4)} \qquad \ln L_1 = -\frac{n}{2} \ln \{\frac{1}{n} [\theta \sum_1 (y_i-\mu)^2 + (1-\theta) \sum_2 (y_i-\mu)^2]\}$$

$$+ \frac{n_1}{2} \ln \theta + \frac{n_2}{2} \ln (1-\theta),$$

corresponding to (5) in the text.

Maximizing (B.4) for θ, we obtain

$$\text{(B.5)} \qquad \hat{\theta} = \frac{n_2^{-1} \sum_2 (y_i-\mu)^2}{n_2^{-1} \sum_2 (y_i-\mu)^2 + n_1^{-1} \sum_1 (y_i-\mu)^2}$$

Finally, inserting (B.5) into (B.4), we obtain the concentrated log likelihood function in μ:

$$\text{(B.6)} \qquad -\ln L_2 = \frac{n_1}{2} \ln \frac{\sum_1 (y_i-\mu)^2}{n_1} + \frac{n_2}{2} \ln \frac{\sum_2 (y_i-\mu)^2}{n_2}$$

Thus, the maximum likelihood estimate of μ, $\hat{\mu}$, is that value of μ which minimizes

$$\text{(B.7)} \qquad Q = \frac{n_1}{n} \ln \frac{\sum_1 (y_i-\mu)^2}{n_1} + \frac{n_2}{n} \ln \frac{\sum_2 (y_i-\mu)^2}{n_2}$$

We now show the consistency of $\hat{\mu}$. We have

$$\text{(B.8)} \qquad \text{plim } Q = (1-P)\cdot\ln E[(y_i-\mu)^2 | y_i \leq \mu] + P\cdot\ln E[(y_i-\mu)^2 | y_i > \mu],$$

where $P = P(y_i > \mu)$. Let us assume $\mu \geqq \mu_0$. Then we have

(B.9) $$P = \int_0^\infty \frac{\sqrt{1-\theta_0}}{\sqrt{2\pi}\sigma_0} \exp\{-\frac{1-\theta_0}{2\sigma_0^2}[z-(\mu_0-\mu)]^2\}dz,$$

(B.10) $$P\cdot E[(y_i-\mu)^2 | y_i > \mu]$$

$$= \int_0^\infty \frac{\sqrt{1-\theta_0}}{\sqrt{2\pi}\sigma_0} z^2 \exp\{-\frac{1-\theta_0}{2\sigma_0^2}[z-(\mu_0-\mu)]^2\}dz$$

$$= \frac{\sigma_0(\mu_0-\mu)}{\sqrt{2\pi(1-\theta_0)}} \exp[-\frac{1-\theta_0}{2\sigma_0^2}(\mu_0-\mu)^2]$$

$$+ \frac{1}{\sqrt{2\pi}}[\frac{\sigma_0^2}{1-\theta_0}+(\mu_0-\mu)^2] \int_{\frac{\sqrt{1-\theta_0}}{\sigma_0}(\mu-\mu_0)}^{\infty} \exp(-\frac{1}{2}z^2)dz$$

and

(B.11) $$P\cdot E[(y_i-\mu)^2 | y_i \le \mu]$$

$$= \int_{-\infty}^{\mu_0-\mu} \frac{\sqrt{\theta_0}}{\sqrt{2\pi}\sigma_0} z^2 \exp\{-\frac{\theta_0}{2\sigma_0^2}[z-(\mu_0-\mu)^2\}dz$$

$$+ \int_{\mu_0-\mu}^{0} \frac{\sqrt{1-\theta_0}}{\sqrt{2\pi}\sigma_0} z^2 \exp\{-\frac{1-\theta_0}{2\sigma_0^2}[z-(\mu_0-\mu)]^2\}dz$$

$$= -\frac{2\sigma_0}{\sqrt{2\pi\theta_0}}(\mu_0-\mu) + \frac{1}{2}[\frac{\sigma_0^2}{\theta_0}+(\mu_0-\mu)^2] + \frac{2\sigma_0}{\sqrt{2\pi(1-\theta_0)}}(\mu_0-\mu)$$

$$- \frac{\sigma_0}{\sqrt{2\pi(1-\theta_0)}}(\mu_0-\mu)\exp[-\frac{1-\theta_0}{2\sigma_0^2}(\mu_0-\mu)^2]$$

$$+ \frac{1}{\sqrt{2\pi}}[\frac{\sigma_0^2}{1-\theta_0}+(\mu_0-\mu)^2] \int_0^{\frac{\sqrt{1-\theta_0}}{\sigma_0}(\mu-\mu_0)} \exp(-\frac{1}{2}z^2)dz$$

From (B.8), (B.9), (B.10), and (B.11) we obtain

(B.12) $$[\frac{\partial}{\partial\mu} \text{plim } Q]_{\mu=\mu_0} = \frac{1}{\sqrt{2\pi}\sigma_0} \{2(\sqrt{\theta_0} - \sqrt{1-\theta_0}) + \sqrt{1-\theta_0}\,[\log (1-\theta_0) - \log \theta_0]\}$$

But the right-hand side of (B.12) can be written as

(B.13) $$\frac{2\sqrt{1-\theta_0}}{\sqrt{2\pi}\sigma_0} (x - \ln x - 1)$$

where $x = \sqrt{\theta_0/(1-\theta_0)}$. Since we have

(B.14) $$\frac{d(x - \ln x - 1)}{dx} \gtreqless 0 \quad \text{as} \quad x \gtreqless 1,\ x \geqq 0,$$

expression (B.13) is nonnegative. Next, consider the case $\mu \leqq \mu_0$. Since

(B.15) $$\text{plim } Q(\theta_0, \mu) = \text{plim } Q(1 - \theta_0,\ 2\mu_0 - \mu),$$

we have

(B.16) $$[\frac{\partial}{\partial\mu} \text{plim } Q]_{\mu=\mu_0} = -\frac{1}{\sqrt{2\pi}\sigma_0} \{2(\sqrt{1-\theta_0} - \sqrt{\theta_0}) + \sqrt{\theta_0}\,[\ln \theta_0 - \ln (1-\theta_0)]\}.$$

But the right-hand side of (B.16) is nonpositive for the same reason that (B.12) is nonnegative. Thus we have shown that plim Q attains a minimum at $\mu = \mu_0$.

It is easy to show from (B.3) and (B.5) that the maximum likelihood estimators of σ^2 and θ are consistent if a consistent estimator of μ is inserted for μ in the right-hand sides of (B.3) and (B.5).

Since the minimum of plim Q is attained at the point where the derivative does not vanish, the asymptotic distribution of the maximum likelihood estimator cannot be obtained by the usual method using a Taylor expansion. Chernoff and Rubin (1956) show that $n(\hat{\mu} - \mu_0)$ has a proper limit distribution but do not obtain it in a closed form. Thus, we are unable to report the proper formulas for the asymptotic standard errors of $\hat{\mu}$, $\hat{\theta}$, and σ^2.

This proof of the consistency of the maximum likelihood estimator carries through directly to the regression case where μ_0 is replaced with $\underline{X}_i'\underline{\beta}_0$ in (B.1), but we do not reproduce that proof here.

The foregoing considerations suggest two potentially useful iterative procedures for determining "pseudo-ML" estimates of parameters.

Method 1. "Corrected least squares".

This method utilizes the first two sample moments of the $\{y_i\}$, namely $\bar{y} = \frac{1}{n}\Sigma_{i=1}^{n} y_i$ and $s^2 = \frac{1}{n-1}\Sigma_{i=1}^{n}(y_i-\bar{y})^2$. We know from (3) and (4) that

$$E(\bar{y}) = \mu + \frac{\sigma}{\sqrt{2\pi}}\left(\frac{\sqrt{\theta} - \sqrt{1-\theta}}{\sqrt{\theta}\sqrt{1-\theta}}\right) \tag{B.17}$$

and

$$E(s^2) = \frac{\sigma^2}{2\theta(1-\theta)}\{1 - \frac{(\sqrt{\theta} - \sqrt{1-\theta})^2}{\pi}\}. \tag{B.18}$$

These two equations suggest the following estimator for μ:

$$\tilde{\mu} = \bar{y} - \frac{s(\sqrt{\theta} - \sqrt{1-\theta})}{[\pi - (\sqrt{\theta} - \sqrt{1-\theta})^2]^{1/2}} \tag{B.19}$$

which is identical to (13) in the text.

The suggested iteration begins by inserting the sample median, which is a consistent estimator of μ, into the right-hand side of (B.5). Second, insert the estimate of θ thus obtained into the right-hand side of (B.19) to obtain the second-round estimate of μ. Repeat the iteration until it converges. Finally, insert the converged values of μ and θ into the right-hand side of (B.3) to obtain an estimate of σ^2.

Using (B.9), (B.10), and (B.11), we can show

$$\text{(B.20)} \qquad \left[\frac{\partial}{\partial\mu} \operatorname{plim} \hat{\theta}\right]_{\mu=\mu_0} = -\frac{4\theta_0^2(1-\theta_0)}{\sigma_0\sqrt{2\pi\theta_0}}$$

and

$$\text{(B.21)} \qquad \left[\frac{\partial}{\partial\mu} \operatorname{plim} \tilde{\mu}\right]_{\theta=\theta_0} = \frac{\sigma_0(\sqrt{\theta_0} - \sqrt{1-\theta_0})}{2\sqrt{2\pi}\theta_0(1-\theta_0)}$$

Since the product of (B.20) and (B.21) is less than 1 in absolute value, the suggested iteration using (B.5) and (B.19) is asymptotically stable in the neighborhood of the true values.

Method 2. "Minimum Distance".

The second method is based on the minimum distance estimator of μ that minimizes

$$\text{(B.22)} \qquad \tilde{S} = \frac{1}{n}\left[\sqrt{\theta}\sum_1 (y_i-\mu)^2 + \sqrt{1-\theta}\sum_2 (y_i-\mu)^2\right]$$

for a given value of θ. The suggested iteration is as follows: First, substitute the sample median for μ in the right-hand side of (B.5). Second, insert the estimate of θ thus obtained into the right-hand side of (B.22) and minimize $\tilde{S}$ with respect to μ to obtain the second-round estimate of μ. Repeat the iteration until it converges. $\tilde{S}$ is continuous in μ and between the adjacent values of y it is smooth. Therefore, it is easy to search for a minimum in the neighborhood of the value of μ obtained in the preceding round of the iteration.

Let $\tilde{S}_0$ be the value of $\tilde{S}$ evaluated at $\theta = \theta_0$ and define $R = \frac{\sqrt{2\pi}}{\sigma_0^2}\tilde{S}_0$. Then we have

(B.23) $$\text{plim } R = \sqrt{\theta_0}\,(1-P)E[(y_i-\mu)^2 \mid y_i \leq \mu] + \sqrt{1-\theta_0}\,PE[(y_i-\mu)^2 \mid y_i > \mu]$$

Assuming $\mu \geqq \mu_0$ and using (B.9), (B.10), and (B.11), we have

(B.24) $$\text{plim } R = -w \exp\left[-\tfrac{1}{2}(1-\theta_0)w^2\right]$$

$$+ \left[\frac{1}{\sqrt{1-\theta_0}} + \sqrt{1-\theta_0}\,w^2\right] \int_{\sqrt{1-\theta_0}\,w}^{\infty} \exp\left(-\tfrac{1}{2}z^2\right)dz + 2w$$

$$+ \frac{\sqrt{2\pi}}{2\sqrt{\theta_0}} + \frac{\sqrt{2\pi\theta_0}}{2}w^2 - \frac{2\sqrt{\theta_0}}{\sqrt{1-\theta_0}}w + \frac{\sqrt{\theta_0}}{\sqrt{1-\theta_0}}w \exp\left[-\frac{1-\theta_0}{2}w^2\right]$$

$$+ \sqrt{\theta_0}\left[\frac{1}{1-\theta_0} + w^2\right] \int_0^{\sqrt{1-\theta_0}\,w} \exp\left(-\tfrac{1}{2}z^2\right)dz$$

Therefore,

(B.25) $$\left|\frac{\partial}{\partial w}\text{plim } R\right|_{w=0} = 0$$

and

(B.26) $$\left|\frac{\partial^2}{\partial w^2}\text{plim } R\right|_{w=0} = \sqrt{2\pi}\left(\sqrt{\theta_0} + \sqrt{1-\theta_0}\right) > 0$$

Therefore, the minimum distance estimator of μ that minimizes $\tilde{S}_0$ is consistent.

It is interesting to note that a similar argument applied to the distance function

(B.27) $$S = \theta \sum_1 (y_i-\mu)^2 + (1-\theta)\sum_2 (y_i-\mu)^2,$$

which is the quadratic term in the concentrated log-likelihood (B.4), results in the conclusion that the minimum distance estimator of μ which minimizes (B.27) at $\theta = \theta_0$ is inconsistent.

These two methods are easily adapted to the regression framework (although the first is by far the easiest to implement). When we take as our goal consistent estimation of the intercept, using the OLS estimators for slope coefficients, then, denoting by $\underline{b}_2$ the $(k-1) \times 1$ vector of OLS slope estimates and by $\underline{X}_{2i}'$ the corresponding ith row of X excluding the first element, one initial consistent estimate of β_1 is just the sample median of $\{y_i - \underline{X}_{2i}'\underline{b}_2\}$.* This initial estimate can be improved upon iteratively along with estimation of θ -- through (B.19) and (B.5) -- or the complete $\underline{\beta}$-vector can be estimated iteratively along with θ through (B.22) and (B.5).

---

*Another consistent estimate comes from applying the MAD estimator to estimate $\underline{\beta}$, which yields an unbiased estimate of the median residual.

Appendix C: Proof that OLS is optimal for the classical regression model when residuals are differentiated according to sign.

When signs of residuals are recognized, the least squares problem is written

(C.1)
$$\min_{\{\underline{\beta},\underline{\varepsilon}\}} \sum_{\varepsilon_i \le 0} \varepsilon_i^2 + \sum_{\varepsilon_i > 0} \varepsilon_i^2$$

$$\text{s.t. } \underline{y} = \underline{X}\,\underline{\beta} + \underline{\varepsilon}.$$

As we have shown in Appendix A, the above can be equivalently written

(C.2)
$$\min_{\{\underline{\beta},\underline{\varepsilon}^+,\underline{\varepsilon}^-\}} \Sigma_{i=1}^n (-\varepsilon_i^-)^2 + \Sigma_{i=1}^n (\varepsilon_i^+)^2$$

$$\text{s.t. } \underline{y} = \underline{X}\,\underline{\beta} + \underline{\varepsilon}^+ - \underline{\varepsilon}^-$$

$$\underline{\varepsilon}^+, \underline{\varepsilon}^- \geq 0$$

$$\min(\varepsilon_i^+, \varepsilon_i^-) = 0 \text{ for all } i,$$

the latter condition being non-binding (i.e. implied by the problem formulation).

By substituting the model identity $\underline{\varepsilon}^+ = \underline{y} - \underline{X}\,\underline{\beta} + \underline{\varepsilon}^-$, we have

(C.3)
$$\min_{\{\underline{\beta},\underline{\varepsilon}^-\}} S^* = \Sigma_{i=1}^n (-\varepsilon_i^-)^2 + \Sigma_{i=1}^n (y_i - \underline{X}_i'\underline{\beta} + \varepsilon_i^-)^2$$

$$\text{s.t. } \varepsilon_i^- \geq 0 \text{ all } i$$

Any solution to this problem must meet the Kuhn-Tucker conditions:

(C.4a) $$\left.\frac{\partial S^*}{\partial \varepsilon_i^-}\right|_{\varepsilon_i^- = \hat{\varepsilon}_i^-} \geq 0 \qquad i=1, \ldots n$$

(C.4b) $$\Sigma_{i=1}^{n}\left[\left.\frac{\partial S^*}{\partial \varepsilon_i^-}\right|_{\varepsilon_i^- = \hat{\varepsilon}_i^-}\right]\hat{\varepsilon}_i^- = 0$$

(C.4c) $$\left.\frac{\partial S^*}{\partial \beta_j}\right|_{\beta_j = \hat{\beta}_j} \geq 0 \qquad j=1, \ldots k$$

(C.4d) $$\Sigma_{j=1}^{k}\left[\left.\frac{\partial S^*}{\partial \beta_j}\right|_{\beta_j = \hat{\beta}_j}\right]\hat{\beta}_j = 0.$$

where $\hat{\varepsilon}_i^-$ (i=1, ... n) and $\hat{\beta}_j$ (j=1, ... k) are solution values for ε_i^- and β_j respectively.

OLS will be optimal if, when $\hat{\underline{\beta}} = (\underline{X}'\underline{X})^{-1}\underline{X}'\underline{y}$, $\hat{\underline{\varepsilon}} = \underline{M}\underline{y}$ with $\underline{M} = I - \underline{X}(\underline{X}'\underline{X})^{-1}\underline{X}'$, and the separation of $\hat{\underline{\varepsilon}}$ into $\hat{\underline{\varepsilon}}^-$ and $\hat{\underline{\varepsilon}}^+$ is <u>ex post</u>, the Kuhn-Tucker conditions are satisfied.

Developing the needed expressions for (C.4a) through (C.4d) for the criterion function (C.3),

(C.5a) $$\frac{\partial S^*}{\partial \varepsilon_i^-} = 2\Sigma_{i=1}^{n} (y_i - \underline{X}_i' \underline{\beta} + \varepsilon_i^-) - 2\Sigma_{i=1}^{n} \varepsilon_i^- , \qquad i=1, \ldots n$$

and

(C.5b) $$\frac{\partial S^*}{\partial \beta_j} = - 2\Sigma_{i=1}^{n} (y_i - \underline{X}_i' \underline{\beta} + \varepsilon_i^-) x_{ij} + $$

$$+ 2\Sigma_{i=1}^{n} (-y_i + \underline{X}_i' \underline{\beta} + \varepsilon_i^+) x_{ij}, \qquad j=1, \ldots k.$$

Now, (C.4a) holds for the OLS solution because $\underline{\iota}'\,\hat{\underline{\varepsilon}} = 0$ by construction, where $\underline{\iota}$ is the $n \times 1$ vector of unities. To see this, when $\hat{\varepsilon}_i \leq 0$, $\hat{\varepsilon}_i^- = -\hat{\varepsilon}_i \geq 0$ and $\hat{\varepsilon}_i^+ = 0$. (C.5a) effectively separates the OLS calculated residuals as to sign and then sums over all of them, first assigning the proper negative sign to otherwise positive $\hat{\varepsilon}_i^-$'s. Since, from (C.4a), $\partial S/\partial\varepsilon_i^- = 0$ for all i when evaluated at $\varepsilon_i^- = \hat{\varepsilon}_i^-$, (C.4b) follows trivially.

The analysis of (C.4c) and (C.4d) follows in much the same way. Since for the OLS residuals, $\hat{\underline{\varepsilon}}'\underline{X}_j = 0$, j=1, ... k, where $\underline{X}_j$ is the j^{th} column of $\underline{X}$, (C.5b) is zero for all j when evaluated at the OLS vectors $\hat{\underline{\varepsilon}}^+$ and $\hat{\underline{\varepsilon}}^-$. (C.4d) is therefore trivially true.

Finally, the qualitative information embodied in recognition of residual signs when signal residuals are weighted equally has no influence on the ML solution either. Referring to text equation (5), we see that when $\theta = 1/2$ the ML criterion function essentially reduces to (C.1).

Again, we caution the reader to the fact that in our model residual signs are not used as a priori knowledge. Obviously, in that instance OLS would be optimal only by coincidence, when the (unconstrained) OLS solution happened to satisfy all the given sign constraints.

REFERENCES

Afriat, S. N., 1972, "Efficiency Estimation of Production Functions," International Economic Review, 13, pp. 568-98.

Aigner, D. J., and S. F. Chu, 1968, "On Estimating the Industry Production Function," American Economic Review, 58, pp. 826-39.

Amemiya, T., 1973, "Regression Analysis when the Dependent Variable is Truncated Normal," Econometrica, 41, pp. 997-1016.

Chernoff, H. and H. Rubin, 1956, "The Estimation of the Location of a Discontinuity in Density," in Proceedings of the Third Berkeley Symposium, Vol. I (J. Neyman, ed.), Berkeley, University of California Press, pp. 19-37.

Jenkins, O. C., et al., 1973, "Root Estimators," Journal of the American Statistical Association, 68, pp. 414-19.

Kendall, M. G., and A. Stuart, 1961, The Advanced Theory of Statistics, Volume 2, New York: Hafner Publishing Co.

Maddala, G. S., and F. D. Nelson, 1973, "Maximum Likelihood Methods for Markets in Disequilibrium," mimeo, University of Rochester. (To appear in Econometrica.)

Poirier, D. J., 1973, Applications of Spline Functions in Economics, Unpublished Ph.D. dissertation, University of Wisconsin, Madison.

Quandt, R.E., 1972, "A New Approach to Estimating Switching Regressions," Journal of the American Statistical Association, 67, pp. 306-10.

Schmidt, P., 1974, "On the Statistical Estimation of Parametric Frontier Production Functions," mimeo, University of North Carolina. (To appear in Review of Economics and Statistics.)

Sielken, R. L., and H. O. Hartley, 1973, "Two Linear Programming Algorithms for Unbiased Estimation of Linear Models," Journal of the American Statistical Association, 68, pp. 639-41.

Timmer, C. P., 1969, On Measuring Technical Efficiency, Unpublished Ph.D. dissertation, Harvard University.

__________, 1971, "Using a Probabilistic Frontier Production Function to Measure Technical Efficiency," Journal of Political Economy, 79, pp. 776-94.